1G5-1

生命科学分野の日本語言語資源の整備と日本語コンテンツへのリンク

Linguistic Resources In Japanese In The Life Science Domain

山本 泰智[*1]
Yasunori Yamamoto

[*1] ライフサイエンス統合データベースセンター
Database Center for Life Science

Database Center for Life Science (DBCLS) has published review articles in Japanese written by Japanese authors who are coauthors of articles published in so called top journals such as Nature or Science. The online review journal is called "Shin-chaku Rombun Review", which means "The latest paper reviews" in English. More than 770 articles have been published and counting. In this situation, it is difficult to find relevant articles to your research interests especially if you are unfamiliar with a field that you search for. To solve this issue, I have developed a navigation system of the review articles by which you can narrow down articles using the MeSH concept tree. Moreover, domain specific terms are identified with Life Science Dictionary (LSD). All of the data behind the system are expressed in Resource Description Framework (RDF) and you can access them using SPARQL in addition to downloading the entire dataset.

1. はじめに

生命科学分野における Nature や Science、Cell などのトップジャーナルといわれる雑誌に発表される論文のうち、日本人が著者に含まれるものを対象に、レビュー論文を当該著者により日本語の書き下ろしで発表する「新着論文レビュー」をライフサイエンス統合データベースセンターではオンラインジャーナルとして発行している[飯田 2013]. 全ての記事が図を含めてオープンであり、クリエイティブコモンズ・ライセンス(CC)表示 2.1 日本を採用している. 各記事は生命科学分野が専門の編集者により綿密に編集作業が行われていることから、良質な当該分野の日本語コンテンツとして多くの研究者に読まれている. 2015 年 3 月 23 日時点で 770 を超す記事が閲覧可能になっており、免疫や発生、シグナル伝達など様々な研究課題に関する記事が含まれている.

そこで筆者は効率良く目的の記事が見つかるような仕組みを提供することで、更に多くの研究者に読まれるようになることを目指し、日本語による概念構造に基づく閲覧システム(提案システム)を開発している. これは新着論文レビューの対象読者として生命科学における専門分野が異なる人を意識していることも背景にあり、見つけたい記事が自身の専門分野ではない場合には適切な検索語を思い浮かばないことも多いと考えられるからである. また、体系化された概念構造を利用した絞り込み機能は新着論文レビューのように広範な生命科学研究全般を扱う場合において有効に働くことがあるという研究[Kashyap 2011]から、開発システムが有益であると想定している.

概念構造の提示に加え、提案システムでは非専門の研究者が様々な概念を元に絞り込みを行う際に、文中で共に使われている用言を確認しやすいように、予め全ての文を対象とした係り受け解析を行い、その結果についても RDF を用いてトリプルストアに格納している. これにより様々な概念について記事の中で記述されている他の概念との関係性を効率的に知ることが出来る. たとえば、アポトーシスという概念については、それが誘導、亢進、促進、制御、抑制などの対象とされることが記事中で述べられていることが容易に分かり、研究対象として生体中での振舞いが多く観察されている現象であると想像できる. また、当該概念を知る研究者は、現在新着論文レビューに含まれる記事で特定の振る舞いについて記述しているものを効率良く見つけることが出来る. すなわち、アポトーシスを抑制することに触れている記事を絞り込むといった使い方が出来る.

提案システムで利用するデータを RDF で表現することの利点として、構築システムの扱うデータの拡張が容易になることがあげられる. また、Linked Data の原則[1]に従う形でオープンライセンスの下に公開することで、構築データの人や機械による再利用性を高めることができる. これらの点についても他に同等のシステムは存在しないことから、提案システムの開発が有益であると認識している.

連絡先:山本泰智, ライフサイエンス統合データベースセンター, 〒277-0871 千葉県柏市若柴 178-4-4, Tel: 04-7135-5508, Fax: 04-7135-5534, yy@dbcls.rois.ac.jp

2. データと構築方法

2.1 データ

提案システムでは生命科学分野で広く知られている Medical Subject Headings (MeSH)[Nelson 2010]という文献検索を効率化するために用意されている英語のシソーラスと、同じく生命科学分野の専門用語について日英対訳や MeSH との関連を収めているライフサイエンス辞書(LSD)[金子 2005]を利用している. これらの特徴や有益性は以前筆者が報告した通りである[山本 2014]が、最新の提案システムでは、リンク対象として、既に RDF データが公開されている日本語 WordNet[Bond 2012]も含めている.

2.2 構築手法

係り受け解析には Cabocha[Kudo 2003]を用いており、その結果を元にして各用言についてそれにかかる格の情報を、助詞部分とその他に分けて RDF を用いたデータを構築している. また出現位置として各文に対する URI を用意して文と用言に関する情報を結びつけている.

これらの情報を表現するための語彙は見つからなかったので独自に開発したオントロジーを用いている (図 1 参照). すなわち、全ての用言と助詞はそれぞれ:Yogen クラスおよび:joshi クラスのインスタンスとし、両クラスはそれぞれ rdfs:Class および

[1] http://www.w3.org/DesignIssues/LinkedData.html

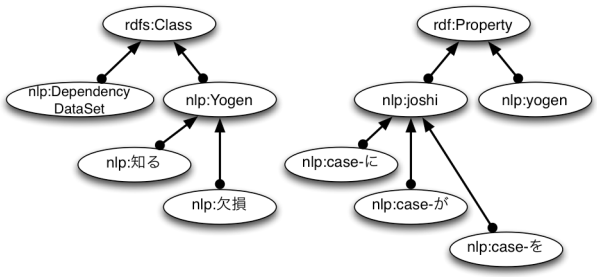

図 1 構築したオントロジー

rdf:Property のサブクラス、サブプロパティとしている. また、文中に出現する各用言とそれに係る格のデータセットを示す URI に対して、:yogen プロパティを用いて当該文中に出現する用言を表し、それに係る句については、句毎に対応する助詞 URI を述語に、残りの部分をそのリテラルの目的語として表現している. なお、用言とそれに係る格のデータセットは:DependencyDataSet クラスのインスタンスとして表現され、更に、dcterms:isPartOf プロパティを用いてそれらの語が出現する文の URI と結ばれている.

例えば、「心筋細胞の分化および発生の過程におけるミトコンドリア融合タンパク質の役割を知るため, マウスの胚において MFN1 および MFN2 を心筋に特異的に欠損させた.」[笠原 2013]という文に対しては、そこには「欠損」と「知る」という二つの用言が含まれているため、それぞれについて RDF を用いたデータセット、すなわち:dependencyDataSet のインスタンスを記述することになる. そして抑制については、以下のようなデータが生成される. なお、ここでは Turtle 形式に準じて表記する.

```
<http://navi.first.lifesciencedb.jp/nlp/7771#1/stc1-1>
  a :dependencyDataSet ;
  :yogen :欠損 ;
  :case-に "特異的" ;
  :case-に "心筋" ;
  :case-において "胚" .
```

図 2 は上述の文に対して得られる RDF グラフの一部である. このようにして得られた係り受けデータと、既に構築されている LSD に含まれている語との関係は SPARQL を用いて結びつけている. すなわち、LSD に含まれている語が出現する文を示す URI を用いて、当該文における係り受け構造のデータを対象とし、各用言に係る句の中に一致するものの有無をみている. 図 2 においては、例えば、「役割」や「特異的」といった語が LSD の語と係り受け解析結果の区の双方に含まれているので、これらが結びつけられる.

3. 提案システム

構築されたデータを用いた提案システム[1]の利用者は最初に MeSH ツリーの最上位階層にある概念を選択する. 以降、特定の概念を選択した時点で当該概念およびその下位概念に含まれる LSD の語が出現する記事の一覧が表示される. 初期設定では語の出現頻度順に記事が並ぶが、出現する語のアルファベット順や記事のタイトル順のほか、元論文の雑誌名や、PubMed ID での並び替えもできる. また、絞り込みを進めた場合に、現在注目している概念の MeSH ツリー階層における位置が表示され、最上位階層からの絞り込みの経緯を容易に知ることが出来る.

また、記事一覧が表示されているページにおいて、出現する語と共に表示されている「Rel」の文字を選択することで係り受け情報を確認できる. そこでは、選択された語について、それが直接係る用言と、それらが出現する文を確認できる. 文を選択することで当該文について RDF で表現されている情報を得ることが出来る. すなわち、それが属する章や他の係り受け情報などである.

4. 結果

執筆時点の 2015 年 3 月 24 日において記事数は 775 あり、全体で 55,758 文が含まれている. なお、文には記事や各章のタイトルも含まれる. また、LSD および日本語 WordNet にリンクしている語の延べ数は 1,507,035、異なり語数は 25,483 である. 延べ数では LSD が全体の 53.4% (805,399) であるのに対し、異なり語数では 63.8% (16,246) である. これは、生命科学分野の語が豊富である LSD に対して、一般的な語の割合が高く、広く使われるが、語数は少ない日本語 WordNet という特徴が適切に反映された結果であると考えられる.

[1] http://navi.first.lifesciencedb.jp/

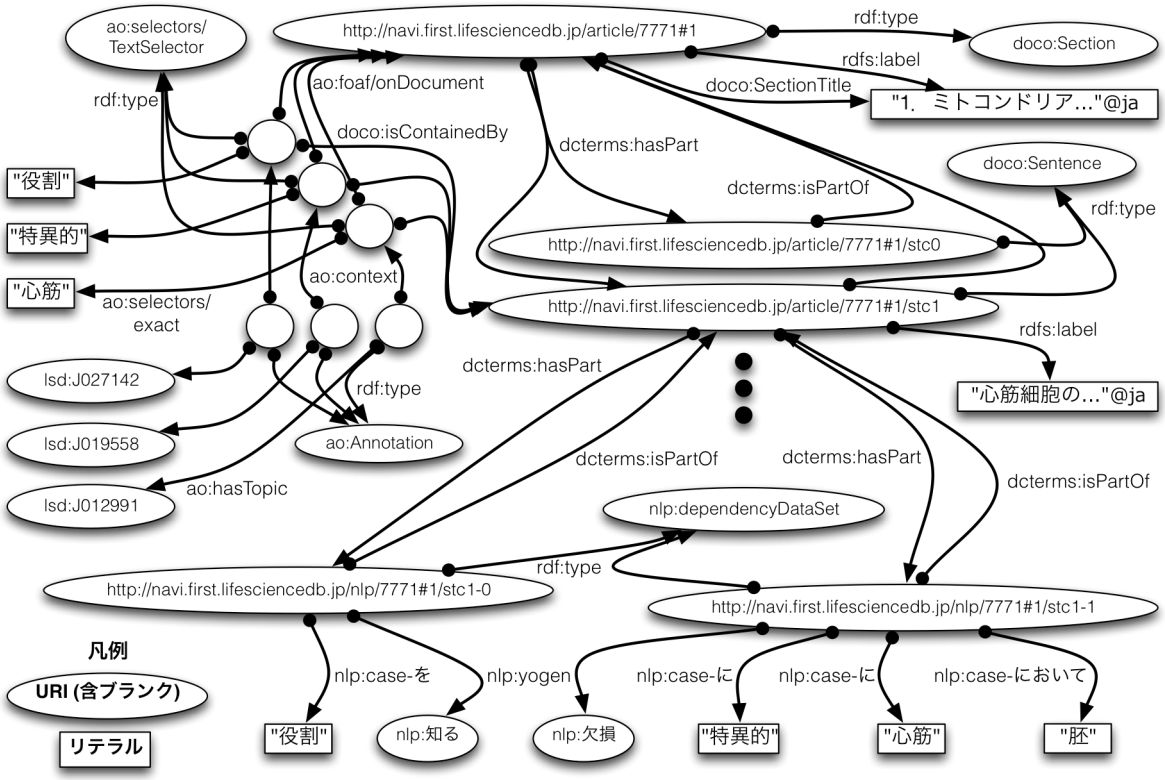

図 2 生成された RDF グラフの一部

また、係り受けデータ数は 178,807、助詞の延べ数は 328,926 である. 従って、用言当たり平均 1.84 の直接係る助詞があることになる. ただ、直接係る助詞の数が 1 であるデータが全体の 44.2% (79,173) であり、最も多い.

構築したデータは Virtuoso (オープンソース版、VOS 7.2.0) に格納している. データにアクセスするためのインターフェースは、ライフサイエンス統合データベースセンターで開発されている Ruby ベースの SPARQL 対応フレームワーク TogoStanza[1] を利用して構築したほか、公開用 SPARQL エンドポイントとして Lodestar と呼ばれる European Bioinformatics Institute (EBI) で開発された Linked Data ブラウザ[2] を利用している.

5. 考察

今後の課題は係り受け解析の精度向上であり、これは、利用している Cabocha のモデルが、新着論文レビューのような生命科学分野の文書を対象として構築されていないことが理由の一つとしてあげられる. このため、今後は新着論文レビューに対するマニュアルでのアノテーション作業を行い、新たなモデルの構築を目指す. 新着論文レビューは一人の編集の専門家による綿密な編集作業を経ていることから、比較的精度の高いモデルの構築がしやすいのではないかと考えている.

また、LSD や日本語 WordNet といった辞書と係り受け解析結果との間のリンクが現時点では SPARQL による文字列マッチでしか行われておらず、このため、係り受け解析結果における句と辞書中の見出し語の間で文字単位のずれが生じていることもある. 特に生命科学分野において重要な遺伝子名やタンパク質名などのアルファベットと数字からなる固有名詞については取りこぼしている例が目立つので対策が必要である.

現時点では日本語 WordNet とのリンクについてはそこに含まれる様々な概念関係を利用しておらず、表層的なリテラルレベルでのリンクにとどまっている. 今後は係り受け解析結果と併せて概念レベルでの情報とリンクするため、各見出し語の曖昧性解消が課題となる.

6. 結論

DBCLS において提供している生命科学分野の最新研究に関する日本語のレビュー、新着論文レビューを対象とした、MeSH 概念階層を利用した記事閲覧システムを構築した. 非専門領域のレビューを効率良く見つけられるように、検索語を必要とせず、MeSH の階層を辿りながら興味のある領域に絞り込むことが可能であり、更に、絞り込まれた研究領域における様々な概念について、記事中で直接係る用言の情報を効率良く得られることから、当該研究領域における知見を学ぶ際に有効に機能すると考えられる.

構築した新着論文レビューの RDF データセットは新着論文レビューと同じくクリエイティブコモンズ・ライセンス(CC)表示 2.1 日本の下、ftp://ftp.dbcls.jp/afara/ で公開しているほか、記事の検索は http://navi.first.lifesciencedb.jp/stanza/top から行える.

7. 謝辞

本研究は独立行政法人科学技術振興機構(JST)、バイオサイエンスデータベースセンター (NBDC) の助成による.

[1] http://wiki.lifesciencedb.jp/mw/BH13.13/TogoStanza
[2] http://www.ebi.ac.uk/fgpt/sw/lodestar/

参考文献

[飯田 2013] 飯田 啓介: 新しい日本語 Web コンテンツ,「新着論文レビュー」と「領域融合レビュー」, 情報管理, Vol. 56, No. 3, pp. 148‐155, 2013.

[Kashyap 2011] Kashyap, A., Hristidis, V., Petropoulos, M., Tavoulari, S.: Effective Navigation of Query Results Based on Concept Hierarchies. Knowledge and Data Engineering, IEEE Transactions on, 23(4), 540-553, 2011.

[Nelson 2010] Stuart J. Nelson, Jacque-Lynne Schulman: Orthopaedic Literature and MeSH, Clinical orthopaedics and related research, Springer, 468(10):2621-2626 , 2010.

[金子 2005] 金子周司, 鵜川義弘, 大武博, 河本健, 竹内浩昭, 竹腰正隆, 藤田信之: ライフサイエンス辞書から生命科学オントロジーへ, 情報知識学会誌, Vol. 15, No. 4, pp. 1‐10, 2005.

[山本 2014] 山本泰智: RDF 化した MeSH とライフサイエンス辞書を利用した生命科学概念に基づく日本語レビュー記事の絞り込み検索, 第 33 回セマンティックウェブとオントロジー研究会, 2014.

[Bond 2012] Bond, F., Baldwin, T., Fothergill, R., Uchimoto, K: Japanese SemCor: A Sense-tagged Corpus of Japanese, The 6th International Conference of the Global WordNet Association (GWC-2012), 2012.

[Kudo 2003] Kudo, T., Matsumoto, Y.: Fast Methods for Kernel-Based Text Analysis, ACL 2003, 2003.

[笠原 2013] 笠原敦, Luca Scorrano: ミトコンドリアの融合は心筋細胞の分化および発生においてカルシニューリンと Notch シグナル伝達系を介し必須である, ライフサイエンス新着論文レビュー, 2013.